Meta-computerlexikografische Bemerkungen zu Vernetzungen in XML-basierten Onlinewörterbüchern – am Beispiel von *elexiko*

Peter Meyer, Institut für Deutsche Sprache Mannheim

Abstract

This contribution outlines a conceptual analysis of the dictionary-internal cross-reference structure in electronic dictionaries along the lines of Wiegand's actional-theoretical text theory of print dictionaries. The discussion focuses on issues of XML-based data modeling, using the monolingual German online dictionary *elexiko* as a running example. The first part of the article demonstrates how Wiegand's formal theory of mediostructure and its intricate nomenclature can be extended in a systematic and lexicographically justified way to cover the structure of the underlying lexicographical database of online dictionaries. The second part of the article applies the concepts developed to a more technical question, examining the extent to which cross-reference information can be stored and processed separately from the dictionary entry documents, e.g., in a relational database. The results are largely negative; in most real world cases, this leads to an unwanted duplication of XML-related structural information. The concluding third part briefly describes the strategy chosen for *elexiko*: mediostructural information is not externalized at all; cross-reference consistency checks are performed by a dictionary editing tool that takes advantage of a specialized XML database index and can easily be made more efficient and scalable by using a simple caching technique.

1. Repräsentation von Verweisstrukturen in elektronischen Wörterbüchern

1.1 Fragestellungen

Konzeptuelle Überlegungen zur Verweis- oder Mediostruktur elektronischer Wörterbücher gestalten sich deutlich vielschichtiger als bei gedruckten Wörterbüchern, weil außer den – prototypisch durch Hyperlinks realisierten – Verweisen auf der Ebene der Präsentation die zugrundeliegende Ebene der Datenmodellierung betrachtet werden muss, die zu Präsentationsaspekten in einem durchaus komplexen Wechselverhältnis steht (vgl. Blumenthal/ Lemnitzer/Storrer 1988). Müller-Spitzer (2007a, 2007b) entwickelt für die konzeptionelle Datenmodellierung von XML-basierten elektronischen Wörterbüchern einen begrifflichen Rahmen analog zu Wiegands (1996, 2002) Theorie der Mediostruktur von Printwörterbüchern. Die vorliegenden Ausführungen schließen in vielerlei Hinsicht an die genannten Arbeiten an und erweitern sie um einige Aspekte mit besonderer Relevanz für die computerlexikografische Praxis. In Abschnitt 1 wird in Ansätzen eine begriffliche Analyse von Vernetzungen, d.h. Verweisstrukturen auf der Ebene der Datenmodellierung, versucht; Abschnitt 2 untersucht dann, inwieweit die Mediostruktur im XML-basierten Wörterbuch in eigenständige Datenstrukturen ausgelagert werden kann.

Die Diskussion wird am Beispiel von *elexiko*, einem korpusbasierten monolingualen Onlinewörterbuch des Gegenwartsdeutschen, geführt, das am Institut für Deutsche Sprache entwickelt wird und unter www.elexiko.de frei zugänglich ist (Haß (Hg.) 2005; Klosa (Hg.) 2011); die grundsätzlichen Überlegungen sind jedoch ohne Weiteres auf andere XML-basierte Wörterbücher übertragbar. Um die Lesbarkeit zu erhöhen, wird in diesem Text zu Beispielzwecken ein deutlich vereinfachtes XML-Schema verwendet.

1.2 Präsentationsebene: Zur begrifflichen Analyse von Verweisen

Die Artikel des Wörterbuchs *elexiko* sind auf der Ebene der Datenbasis *inhaltsorientiert* ausgezeichnet, vergleichbar dem *lexical view* in den aktuellen TEI-Richtlinien[1]; jedem Artikel des Wörterbuchs entspricht – und diese Voraussetzung gilt für den gesamten vorliegenden Beitrag – ein separates XML-Dokument, dessen *Inhaltsstruktur* durch die DTD bzw. das XML-Schema des Wörterbuchs vorgegeben ist (zu den Begrifflichkeiten vgl. im Einzelnen Müller-Spitzer 2007b). Wir betrachten im Folgenden den typischen Fall von Verweisangaben zwischen *elexiko*-Artikeln und greifen das Beispiel paradigmatischer Relationen (Sinnbeziehungen wie Synonymie, Hyponymie etc.) zwischen Lesarten verschiedener Lemmata heraus. Auf der Präsentationsebene erscheinen die Verweise als Hyperlinks auf einer lesartspezifischen Registerkarte für sinnverwandte Wörter (vgl. Abb. 1). In der in Wiegand (2002) für das Printmedium entwickelten, begrifflich natürlich an die veränderten Bedingungen des Onlinemediums anzupassenden Terminologie ist eine solche *Verweisangabe* artikelintern an eine *Verweisausgangsangabe* (als *Bezugsadresse* der Verweisangabe, die den *Verweisausgangsbereich* identifiziert, von dem aus verwiesen wird) adressiert, die durch das betrachtete Lemma und eine die betrachtete Lesart semantisch umschreibende Kurzetikettierung oberhalb der Registerkarte gebildet wird. In Abbildung 1 ist der Verweisausgangsbereich informell durch „Artikel *Familie*, Lesart ‘Verwandte’“ beschreibbar. Mediostrukturell ist die Verweisangabe an die *Verweisadresse* adressiert, an die der Benutzer verwiesen wird und mit der er wiederum auf den zugehörigen *Verweiszielbereich* (hier ebenfalls durch die Angabe von Lemma und Lesartenkurzetikettierung identifiziert) verwiesen wird. Die Verweisangabe selber steht in einer (in der HTML-Darstellung zweidimensional zu charakterisierenden) *Verweisposition*, die ausdrücklich nicht identisch ist mit dem Verweisausgangsbereich. Die Verweisangabe[2] selber enthält eine *Verweisadressenangabe*, die offenbar nicht in herkömmlicher Weise durch ein Textsegment gegeben ist; klickt man etwa bei den Synonymen zu *Familie* / Lesart ‘Verwandte’ auf das Wort *Haushalt*, wird man per Hyperlink zum *elexiko*-Artikel *Haushalt* in der Lesart ‘Personengruppe’ weitergeleitet; relevant ist hier also das im HTML- / JavaScript-Quellcode der Seite – im typischen Fall durch eine URL-Angabe – spezifizierte Interaktionsverhalten der betreffenden Ansicht des Onlinewörterbuchs, nicht die in diesem Falle verkürzte textuelle Spezifikation, die der Benutzer im Browserfenster sieht.[3]

1.3 Datenmodellierungsebene: Zur begrifflichen Analyse von Vernetzungen

Das soeben geschilderte visuell-interaktionale Verständnis der Präsentationsebene ist begrifflich von den Verhältnissen in der *lexikografischen Datenbasis*, vereinfacht gesagt also der Gesamtheit der den Wortartikeln zugrunde liegenden XML-Dokumente, zu trennen. Diese Trennung hat zum einen naheliegende technische Gründe, denn die tatsächliche Kodierung des Verweiszieles in einem XML-Dokument wird von der Webapplikation auf komplexe und letztlich arbiträre Weise in eine URL-basierte Adressierung von dynamisch generierten Webseiten „übersetzt“, wobei Webseiten (oder, im Falle von AJAX-Anwendungen, auch HTML-

[1] Der *lexical view* wird in den aktuellen Richtlinien (TEI-P5) wie folgt erläutert: „this view includes the underlying information represented in a dictionary, without concern for its exact textual form“ (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html); weitere „views“ sind der *typographic view* und der *editorial view*.

[2] Wir unterscheiden hier nicht zwischen (rein verweisvermittelnden) Verweisangaben und (nicht ausschließlich verweisvermittelnden) Angaben mit Verweiskennzeichnung. Im Text behandeln wir nur erstere; die vorgestellten Überlegungen lassen sich jedoch auf letztere übertragen.

[3] Zu Hyperlinks aus metalexikografischer Sicht vgl. noch Kammerer (1998), dem jedoch ersichtlich das Konzept einer konzeptuellen Datenmodellierung noch nicht zur Verfügung stand.

Fragmente) in keiner simplen Entsprechungsrelation zu lexikografischen Zieladressen stehen. Zum anderen korrespondiert, wie in Müller-Spitzer (2007a, S. 142ff.) ausführlicher erläutert wird, nicht jede Verweisrelation auf der Präsentationsebene mit einer Vernetzungsbeziehung in der Datenbasis und umgekehrt.

Abb. 1: Ausschnitt aus der Onlinepräsentation des *elexiko*-Artikels *Familie* in der Lesart ‘Verwandte’; Angabebereich „Sinnverwandte Wörter“ (http://www.owid.de/artikel/3920/Verwandte; Stand Oktober 2012)

Im Folgenden betrachten wir anhand der paradigmatischen Relationen in *elexiko* den einfachsten Fall, in dem ein bestimmter Typ von Verweisangabe auf der Präsentationsebene tatsächlich systematisch – d.h. letztlich aufgrund der Programmierung der Webapplikation – einem bestimmten Typ von *Vernetzung* auf der Ebene der XML-Datenbasis entspricht. Unter Vernetzung verstehen wir hier zunächst nur den rein formal definierbaren Sachverhalt, dass ein XML-Element z.B. durch seine Attribute ein anderes XML-Element (typischerweise in einem anderen XML-Dokument derselben Datenbasis) referenziert. Die Referenzierung geschieht üblicherweise über ein Adressierungsverfahren, das – für einen bestimmten Typ von

XML-Elementen[4] – eine bijektive (eineindeutige) Zuordnung von Elementen eben dieses Typs zu IDs (identifizierenden Zeichenketten) oder geordneten n-Tupeln von IDs herstellt. Letzterer Fall lässt sich einfacher anhand eines deutlich vereinfachten und schematisierten Ausschnittes aus dem XML-Dokument des *elexiko*-Wortartikels *Familie* erläutern (siehe Abb. 2, der entsprechende Ausschnitt aus der Onlinepräsentation ist in Abb. 1 dargestellt).

```
<artikel id="1234">
  <lemmazeichen>Familie</lemmazeichen>
  …
  <lesart id="Verwandte">
    …
    <paradigmatik>
      …
      <synonymie>
        …
        <relpartner art-id="5678" lesart-id="Personengruppe">
          Haushalt
        </relpartner>
      </synonymie>
      …
    </paradigmatik>
    …
  </lesart>
  …
</artikel>
```

Abb. 2: Vereinfachter Ausschnitt aus dem XML-Dokument des *elexiko*-Wortartikels *Familie*: Synonymievernetzung von der Lesart ‘Verwandte’ auf die Lesart ‘Personengruppe’ des Artikels *Haushalt*. Die für Adressierungen verwendeten IDs sind fett gesetzt.

Der Artikel als Ganzes und damit auch das den gesamten Artikel umfassende Dokumentelement ist durch die ID ‘1234’ in der zugrundeliegenden Datenbasis eineindeutig identifizierbar; die Lesart mit der Kurzetikettierung ‘Verwandte’ ist durch das geordnete Paar (‘1234’, ‘Verwandte’) eineindeutig identifizierbar; innerhalb von Lesarten werden in *elexiko* – im Beispiel nicht gezeigt – gelegentlich noch Lesartenspezifizierungen unterschieden, die dann durch ein geordnetes Tripel von IDs, z.B. (‘1234’, ‘Verwandte’, ‘Dynastie’), identifiziert werden. Die Vernetzung auf das Lemma *Haushalt* in der synonymen Lesart ‘Personengruppe’ geschieht, wie im Beispiel ersichtlich, durch Angabe des zugehörigen geordneten ID-Paares (‘5678’, ‘Personengruppe’); die Angabe der Lemmazeichengestalt *Haushalt* ist redundant und erfolgt aus pragmatischen Gründen, denn so kann die zugehörige HTML-Repräsentation ausschließlich auf der Grundlage dieses einen XML-Dokuments gewonnen werden, im Falle von *elexiko* durch XSL-Transformationen; es müssen nicht erst in der Datenbank sämtliche Lemmazeichenangaben in den adressierten Artikeln nachgeschlagen werden. – Natürlich ist die Identifikation durch geordnete Tupel von IDs, die zu ineinander verschachtelten Elementen gehören, nicht die einzige Möglichkeit eines Adressierungsschemas. Man kann auch jedes zu adressierende Element einer Datenbasis separat mit einer eineindeutigen ID versehen. Die hier demonstrierte Vorgehensweise kann aber konzeptionelle Vorteile haben, wie in Abschnitt 2.4 deutlich werden wird.

[4] Wir vereinfachen hier durchgängig, indem wir in vielen Fällen von Elementen sprechen, wo allgemeiner von Knoten im XML-Dokumentenbaum gesprochen werden müsste. Die erforderliche Verallgemeinerung ist aber trivial und trägt nichts zu den hier diskutierten Aspekten bei.

Bei der Analyse der hier exemplarisch gezeigten Vernetzung ist es durchaus möglich, die oben zur Beschreibung von Verweisen auf der Präsentationsebene verwendeten Wiegand'schen Begrifflichkeiten in ihrem Anwendungsbereich entsprechend „auszudehnen".[5] Eine Übertragung dieser handlungstheoretisch fundierten begrifflichen Analyse auf so etwas wie die Datenmodellierung ist im Fall von *elexiko* schon dadurch gerechtfertigt, dass die *lexikografischen* Handlungen direkt in den XML-Dokumenten vorgenommen werden. Die Artikelautoren arbeiten also direkt mit der XML-Repräsentation des Artikels. Natürlich ist es ihnen möglich, die Perspektive des Benutzers einzunehmen und sich beispielsweise die dem XML-Dokument korrespondierenden Onlineansichten anzuschauen. Diese können sich aber jederzeit ändern, beispielsweise um neue Darstellungsaspekte und Funktionalitäten erweitert werden. Insgesamt wird eine vollständige begriffliche Analyse – die hier nicht geleistet werden soll – durch die Existenz von zwei durch komplexe Regeln miteinander verknüpften „Textebenen" für die Produktion bzw. Nutzung / Rezeption wesentlich komplexer.

Wir zeigen nun am eingeführten Beispiel, wie die oben eingeführten Begriffe bestimmten Aspekten des betrachteten XML-Dokuments aus Abb. 2 zugeordnet werden können; dabei ersetzen wir in den verwendeten Termini jeweils den Wortbestandteil *-verweis-* durch *-vernetzung-*.

- Der zur Vernetzungsangabe gehörende *Vernetzungsausgangsbereich* ist das artikelintern adressierte übergeordnete XML-Element, von dem aus auf ein anderes XML-Element vernetzt wird; da hier eine Synonymievernetzung zwischen Lesarten von zwei Artikeln vorliegt, wäre es vielleicht naheliegend, das XML-Element <lesart>[6] mit diesem Bereich zu identifizieren. Korrekterweise ist jedoch <synonymie> der Vernetzungsausgangsbereich, da ja auf *Haushalt* in der Lesart 'Personengruppe' *als Synonym* und nicht z.B. als Partonym vernetzt (und auf der Präsentationsebene verwiesen) werden soll. Die zum genannten Bereich gehörende *Vernetzungsausgangsangabe* besteht aus der für eine eindeutige Identifikation erforderlichen Eigenschaft (z.B. ID-Attributwert oder Elementname) des betreffenden Elements selber sowie der nach dem verwendeten Bijektionsschema *relevanten*, mit IDs versehenen Vorfahren des Elements. Genauer lässt sich die Vernetzungsausgangsangabe mit der standardisierten XML-Abfragesprache XPath beschreiben; im Beispiel handelt es sich um den XPath-Ausdruck **/artikel[@id='1234']//lesart [@id='Verwandte']//synonymie**, der die gewünschte Einer-Knotenmenge ausdrückt. Die Vernetzungsausgangsangabe spezifiziert die *Bezugsadresse* der Vernetzung, die hier aus dem geordneten ID-Tripel ('1234', 'Verwandte', 'synonymie') besteht.[7]
- Die Vernetzungsangabe selber befindet sich innerhalb des Vernetzungsausgangsbereichs[8] (in XML-Terminologie: als Nachkomme an einer nach dem verwendeten XML-Schema dafür vorgesehenen *Vernetzungsposition*). In diesem Fall handelt es sich einfach um einen von mehreren Kindelementen des Ausgangsbereichs. Die

[5] Nur am Rande sei vermerkt, dass ein solcher Umgang mit Begriffen (nicht bloß Termini) nicht dem grundsätzlichen wissenschaftstheoretischen Verständnis von Begriffsbildungen zuwiderläuft, sondern sogar grundlegende Eigenschaft des „Funktionierens" von Begriffen ist; so argumentiert umfassend und mit zahlreichen Beispielen v.a. aus den Natur- und Ingenieurswissenschaften Wilson (2006).

[6] Die hier verwendete verkürzende Bezeichnungsweise sollte hinreichend klar sein; gemeint ist ein konkretes XML-Element des Beispieldokuments, nämlich hier das einzige im Beispiel gezeigte Element, dessen Tagname gleich *lesart* ist.

[7] Im Allgemeinen müssen IDs also keine XML-Attributwerte sein, es kann sich auch z.B. um Elementnamen handeln. – Hier soll ohne Beschränkung der Allgemeinheit davon ausgegangen werden, dass sich mit den Eigenschaften der IDs (z.B. Zahlenbereich) und der Stelligkeit des ID-Tupels eindeutig das zugehörige XML-Element identifizieren lässt.

[8] Dies ist natürlich keine logisch zwingende Entscheidung; die Angabe könnte z.B. auch ein Geschwisterknoten des Ausgangsbereichsknotens sein.

Vernetzungsangabe selber ist hier das <relpartner>-Element, das die *Vernetzungsadressenangabe* enthält, in diesem Falle in Gestalt von zwei Attributwerten, die die Artikel- und Lesart-IDs des synonymen Relationspartners angeben. Als die von der Angabe spezifizierte *Vernetzungsadresse* fassen wir wiederum das zugehörige geordnete Paar (‘5678’, ‘Personengruppe’) von IDs auf. – Weiter unten besprechen wir den Fall, dass die Vernetzungsangabe keine Vernetzungsadressenangabe enthält, sondern über eine eindeutige ID verfügt, aus der in einer externen Tabelle die Vernetzungsadresse ermittelt werden kann.

- Die genannte Vernetzungsadresse ist die Adresse des *Vernetzungszielbereichs*, der sich im hier nicht gezeigten XML-„Zieldokument“ des Wortartikels *Haushalt* befindet; dabei handelt es sich um das <lesart>-Element dieses Artikels mit dem ID-Attribut ‘Personengruppe’.

2. Wie ‘selbstständig’ ist die Mediostruktur in XML-basierten Wörterbüchern?

2.1 Vernetzungsspezifikationen

Das geordnete Paar aus Bezugs- und Vernetzungsadresse, im Beispiel: ([‘1234’, ‘Verwandte’, ‘synonymie’], [‘5678’, ‘Personengruppe’]), stellt gewissermaßen den informationellen, mediostrukturellen Kern der Vernetzungsbeziehung dar und soll hier als (*lexikografische*) *Vernetzungsspezifikation* bezeichnet werden. Das damit adressierte Paar aus Vernetzungsausgangs- und -zielbereich möge nunmehr auch kurz als (*lexikografische*) *Vernetzung* bezeichnet werden; dies entspricht im Wesentlichen, wenn auch mit abweichender Terminologie, der Definition unidirektionaler Vernetzungen als geordneten Paaren aus adressierten Quell- und Zielressourcen bei Müller-Spitzer (2007a: 167).

Aus computerlexikografischer Sicht von besonderer Relevanz ist die Menge der Vernetzungsspezifikationen eines elektronischen Wörterbuchs, die mengentheoretisch-extensional eine Relation ist. Sie kann als Grundlage für wesentliche Konsistenzprüfungen dienen, die für ein automatisiertes Vernetzungsmanagement von Bedeutung sind und nicht sinnvoll in manueller lexikografischer Arbeit erledigt werden können. Hierbei geht es zum einen darum, ob die Vernetzungsadresse in der Datenbasis überhaupt existiert, zum anderen aber auch, ob die genannte Relation oder Teilmengen davon (z.B. die Menge der Synonymierelationen) bestimmte Bedingungen wie Symmetrie[9] oder Transitivität erfüllen.

Da Vernetzungen in der Regel artikelübergreifend sind, liegt es nahe, bei einer XML-basierten Repräsentation der Artikel die Vernetzungsspezifikationen in einer gesonderten Datenstruktur zu speichern, beispielsweise in einer relationalen Datenbanktabelle, die performante Suchen gestattet. Ganz allgemein ist die Frage, inwieweit sich die Mediostruktur eines XML-basierten elektronischen Wörterbuchs unabhängig von der hierarchischen Struktur der Einzelartikel repräsentieren lässt, von computerlexikografischem Interesse, und soll im Folgenden ausführlicher betrachtet werden.

[9] Die Darstellung ist hier wieder etwas vereinfacht: Wenn man die Vernetzungsspezifikationen z.B. auf Synonymie einschränkt, kann man sie verkürzend notieren, ohne den Elementnamen <synonymie> anzugeben; erst dann kann man bezüglich geordneter Paare der Art ([‘1234’,‘Verwandte’], [‘5678’,‘Personengruppe’]) von Symmetrie der Relation sprechen.

2.2 Vernetzungspositionen und Separabilität der Mediostruktur

Konkret lässt sich zunächst fragen, ob es im Allgemeinen durch eine Reorganisation der XML-Datenbasis eines Wörterbuchs möglich ist, Vernetzungsspezifikationen dergestalt in einer separaten Datenstruktur zu verwalten, dass das Anlegen oder Löschen einer Vernetzung *keine Änderungen am XML-Dokument erforderlich macht*. Diese mögliche Eigenschaft der XML-Datenbasis eines Wörterbuchs sei hier kurz als *Separabilität der Mediostruktur* bezeichnet. Die Antwort auf die gestellte Frage ist negativ; für eine nähere Begründung kommen wir auf den in Wiegands Arbeiten eher unscharf verwendeten Begriff der Verweis- bzw. hier Vernetzungs*position* zurück. In unserem Beispiel sind die Vernetzungsangaben zu den Synonymen zu *Familie* in der Lesart 'Verwandte' eine Menge von Geschwisterknoten, die laut XML-Schema sämtlich Kindelemente des Elements <synonymie> sind und in diesem Sinne eine bestimmte *schemainduzierte generische Vernetzungsposition* gemein haben. Die Reihenfolge der Vernetzungsangaben ist jedoch nicht mehr durch ein Schema vorgebbar. Die *konkrete Vernetzungsposition* der Vernetzungsangaben kann daher im Allgemeinen nicht mechanisch aus der Vernetzungsspezifikation abgeleitet werden. In diesem speziellen Fall ist sie in *elexiko* einfach durch eine alphabetische Sortierung gegeben, so dass es allein aufgrund der Vernetzungsspezifikationen immerhin noch möglich wäre, die konkreten Positionen der Angaben mechanisch zu bestimmen – allerdings nur dann, wenn sämtliche Spezifikationen, die zu einer und derselben schemainduzierten generischen Vernetzungsposition gehören, bekannt sind. Bestünden die Angabebereiche zu den verschiedenen paradigmatischen Relationen (Synonymie, Hyperonymie, …) sämtlich nur aus Verweisen auf alphabetisch geordnete Relationspartner, wäre die Mediostruktur – bzw. zumindest der hier betrachtete Ausschnitt der Mediostruktur – separabel, es wäre sogar möglich, das gesamte Element <paradigmatik> aus dem XML-Schema herauszunehmen und statt dessen die zugehörigen Informationen in einer simplen relationalen Tabelle von Vernetzungsspezifikationen zu speichern.

Der eben eingeführte Begriff der konkreten Vernetzungsposition lässt sich formal genauer fassen: Er wird durch einen XPath-Ausdruck repräsentiert, der auf dem Vernetzungsausgangsbereich operiert und das zur Vernetzungsangabe gehörende XML-Element eindeutig spezifiziert. Diese Präzisierung geht davon aus, dass die Vernetzungsposition nicht durch eine eigene ID erschließbar ist, so dass der XPath-Ausdruck im Allgemeinen mit *Zugriffsindizes* arbeiten muss.[10] In unserem Beispiel aus Abb. 1 / 2 wäre, da *Haushalt* das an dritter Stelle genannte Synonym und der Vernetzungsausgangsbereich das Element <synonymie> ist, die konkrete Vernetzungsposition einfach durch den auf das <synonym>-Element anzuwendenden XPath-Ausdruck **relpartner[3]**, oder, in der ausführlichen XPath-Notation, **child::relpartner[position() = 3]** gegeben.

2.3 Nichtseparable Mediostrukturen I: Die Rolle positionaler Informationen

Wäre die Abfolge der Relationspartner Gegenstand inhaltlich-lexikografischer Entscheidungen und könnte deswegen nicht mechanisch aus den Vernetzungsspezifikationen erschlossen werden, wären die oben gegebenen Bedingungen für Separabilität nicht mehr gegeben; dadurch würde eine getrennte Datenhaltung für die Mediostruktur, ähnlich wie bei den nachfolgenden Beispielen, allerdings nicht unmöglich, aber bereits deutlich komplexer, da man nun beispielsweise den XPath-Zugriffsindex als zusätzliche Information in die Tabelle der Vernetzungsspezifikationen aufnehmen müsste und damit einen Aspekt der Hierarchisierung

[10] Abstrakter formuliert muss der XPath-Ausdruck im Allgemeinen auch entlang der Achsen **preceding** und **following** navigieren.

und Abfolge von Elementen der XML-Struktur zusätzlich in die separate Datenstruktur auslagern müsste. Ein Redaktionssystem müsste dafür sorgen, dass diese Aufteilung der Datenhaltung in der Benutzeroberfläche für den Lexikografen transparent ist, und müsste zudem die Konsistenz der Datenhaltung sicherstellen, etwa dadurch, dass geprüft wird, ob die Zugriffsindizes für zusammengehörige konkrete Vernetzungspositionenen wirklich bei 1 beginnen und fortlaufend sind. Dieses neue Konsistenzproblem wäre ein Artefakt der getrennten Repräsentation von Mikro- und Mediostrukturen.

Noch deutlicher wird die Problematik anhand des in Abbildung 3 gezeigten weiteren Ausschnittes aus der Paradigmatik des Artikels *Familie*, zu dem in Abbildung 4 ein wiederum deutlich vereinfachter Ausschnitt des XML-Dokuments gezeigt wird. In *elexiko* können mehrere Relationspartner, für die ein gemeinsamer Korpusbeleg vorliegt, zu einer <beleg-gruppe> zusammengefasst sein. Überdies können sämtliche Relationspartner bzw. Beleggruppen für eine gegebene Sinnrelation zu mehreren „Relationspartnergruppen“ zusammengefasst sein; so wird hier zwischen Partonymen zu ‘Familie im engeren Sinne’ und zu ‘Familie im weiteren Sinne’ unterschieden. Um diese mehrfache hierarchische Untergliederung nach Gruppen von Vernetzungsangaben in eine separate Datenstruktur auslagern zu können, würde man bereits ein recht ausgefeiltes System von Positions- und/oder Gruppenindizes benötigen, dessen einziger Zweck zudem darin bestünde, Lagebeziehungen zwischen Knoten in XML-Bäumen zu reproduzieren. Man kann in diesem Zusammenhang ein Konzept von *schwacher Separabilität* der Mediostruktur definieren, das genau dann zutrifft, wenn es möglich ist, Vernetzungsspezifikationen *zusammen mit positionalen Indizes für die zugehörigen Vernetzungsangaben* in einer separaten Tabelle so zu speichern, dass das Anlegen oder Löschen einer Vernetzung ohne Veränderung des XML-Dokuments möglich ist.

Abb. 3: Weiterer Ausschnitt aus der Onlinepräsentation des *elexiko*-Artikels *Familie* in der Lesart ‘Verwandte’ (http://www.owid.de/artikel/3920/Verwandte; Stand Oktober 2012), Angabebereich „Sinn-verwandte Wörter“

```
<artikel id="1234">
  <lemmazeichen>Familie</lemmazeichen>
    …
    <lesart id="Verwandte">
      …
      <paradigmatik>
        …
        <partonymie>
          <relpartner-gruppe titel="Familie im engeren Sinne">
            …
            <relpartner art-id=… lesart-id=…>Elternteil</relpartner>
              …
            <beleg-gruppe>
              <relpartner art-id=… lesart-id=…>Mutter</relpartner>
              <relpartner art-id=… lesart-id=…>Tochter</relpartner>
              <beleg> … </beleg>
            </beleg-gruppe>
             …
          </relpartner-gruppe>
          <relpartner-gruppe titel="Familie im weiteren Sinne">
            …
          </relpartner-gruppe>
        </partonymie>
        …
    </paradigmatik>
    …
  </lesart>
  …
</artikel>
```

Abb. 4: Vereinfachter Ausschnitt aus dem XML-Dokument des *elexiko*-Wortartikels ‘Familie’: Vernetzung von der Lesart ‘Verwandte’ auf eine Gruppe von partonymen Relationspartnern. Die gegenüber Abb. 2 neu hinzugekommenen XML-Elemente sind fett gesetzt

2.4 Nichtseparable Mediostrukturen II: Angabezusätze

Eine weitere Komplikation ergibt sich offenbar durch die bislang ignorierten *Angabezusätze* zu Vernetzungsangaben, wie etwa Korpusbelege und weitere Hinweise (vgl. erneut Abb. 1), die sich inhaltlich ausschließlich auf die Vernetzung beziehen. In der hier verwendeten, technischen Definition von Vernetzungsspezifikation sind solche Zusätze nicht berücksichtigt: Wenn man eine lexikografische Datenbasis ohne vernetzungsbezogene Angabezusätze, deren Mediostruktur separabel oder schwach separabel ist, um solche Zusätze anreichert, bleibt die Mediostruktur separabel bzw. schwach separabel. Dies lässt sich an den obigen Beispielen leicht nachvollziehen: Es ist beim Vorhandensein von vernetzungsbezogenen Angabezusätzen zwar nicht mehr möglich, durch Auslagern der Vernetzungsspezifikationen das gesamte <paradigmatik>-Element zu streichen, aber man kann – trivialerweise – immer noch die Vernetzungsadressenangaben (Attribute) aus den <relpartner>-Elementen entfernen. Aus lexikografischer wie informatischer Sicht ist ein solches Vorgehen nicht überzeugend, weil nunmehr vernetzungsbezogene Informationen in wenig einleuchtender Weise auf XML-Dokument und externe Tabelle der Vernetzungsspezifikationen verteilt sind. Inhaltlich gesehen lassen sich Vernetzungen auch als ternäre Beziehungen auffassen – es wird a) von etwas b) mit etwas c) auf etwas vernetzt; die Angabezusätze sind ein Aspekt des „mit etwas“. Man kann daher ein Konzept *erweiterter Vernetzungsspezifikationen* einführen, die geordnete Paare aus einer

Vernetzungsspezifikation sowie einer Spezifikation der ausschließlich zur jeweiligen Vernetzung gehörigen Angabezusätze sind. Letztere müssten in komplexeren Fällen, so wie sie in *elexiko* vorliegen, als XML-Fragment oder auch, falls überhaupt möglich, als komplexes relationales Äquivalent eines solchen Fragments repräsentiert werden, wenn man versuchen möchte, diese Spezifikationen als separate Datenstruktur abzuspeichern. Solange solche Zusätze strikt jeweils genau einer Vernetzungsangabe zugeordnet sind, ist ein solches Auslagern aller vernetzungsbezogenen Informationen in erweiterte Vernetzungsspezifikationen dergestalt möglich, dass wiederum das Anlegen oder Löschen von Vernetzungen einschließlich der Angabezusätze ohne Veränderungen am XML-Dokument möglich ist. Dies lässt sich im hier eingeführten terminologischen Stil als *erweiterte Separabilität* bzw. *erweiterte schwache Separabilität der Mediostruktur* bezeichnen, je nachdem, ob auch positionale Indizes mit den Vernetzungsspezifikationen abgespeichert werden müssen. Die Mediostruktur von *elexiko* ist jedoch nach der eben absichtlich so restriktiv eingeführten Definition nicht einmal erweitert schwach separabel, da Angabezusätze sich häufig auf mehr als eine Vernetzung beziehen, wie sofort aus Abb. 1 und Abb. 3 ersichtlich ist. So ist es gerade das charakteristische Merkmal von Beleggruppen, dass sie einen gemeinsamen Korpusbeleg haben; Vernetzungen in Relationspartnergruppen haben eine gemeinsame erläuternde Überschrift; usw. Natürlich kann man auch solche übergreifenden Angabezusätze in eine separate Struktur auslagern; der Preis dafür wäre dann jedoch, dass man die ausgelagerten Vernetzungsspezifikationen analog den Strukturen des ursprünglichen XML-Dokuments hierarchisch gliedern und die auszulagernden Zusätze dann an Gruppen von Vernetzungsspezifikationen zuweisen müsste – womit die grundsätzliche Idee, die hierarchische interne Gliederungsstruktur eines Wörterbuchartikels von den mediostrukturellen Beziehungen zwischen Artikeln zu trennen, endgültig ad absurdum geführt wäre.

Dort, wo keine hinreichend einfache Form von Separabilität der Mediostruktur gegeben ist, steht nun noch ein anderer Weg offen, um die u.U. gewünschte informationelle Trennung zu erreichen: Man kann zunächst die Vernetzungsangaben selber in das allgemeine Adressierungsschema mit einbeziehen, also durch IDs oder ID-Tupel identifizierbar machen. In einem zweiten Schritt entfernt man dann lediglich die Vernetzungsadressenangaben, in unseren Beispielen die Attribute @art-id und @lesart-id, und lagert die reine Vernetzungsinformation in eine *strukturelle Vernetzungsspezifikation* aus, die einfach ein geordnetes Paar aus der Adresse der Vernetzungsangabe und der Vernetzungsadresse ist. Im Beispiel aus Abb. 1 hätten wir dann statt eines Vernetzungsangabe-Elements **<relpartner art-id="5678" lesart-id="Personengruppe">…</relpartner>** ein Element **<relpartner id="rel_haushalt">…</relpartner>**; in einer separaten Tabelle würde dann die strukturelle Vernetzungsspezifikation (['1234', 'Verwandte', 'synonymie', 'rel_haushalt'], ['5678', 'Personengruppe']) abgespeichert. Die soeben erneut illustrierte Adressierungstechnik mittels geordneter ID-Tupel hat in der hier gezeigten Form den Vorteil, dass eine solche strukturelle Vernetzungsspezifikation immer zugleich auch eine gewöhnliche Vernetzungsspezifikation ist und mithin Auskunft über den Vernetzungsausgangsbereich liefert, was für Konsistenzprüfungen nützlich ist.

Im Ergebnis ist die Auslagerung nur der strukturellen Vernetzungsspezifikationen eine in allen Fällen verfügbare, computerlexikografisch recht saubere Lösung, denn sie lässt die hierarchische, XML-basierte Datenmodellierung innerhalb der Wortartikel intakt und separiert nur dasjenige Minimum an artikelübergreifender mediostruktureller Information, das für Korrektheits- und Konsistenzprüfungen benötigt wird, nämlich die Vernetzungsadresse und ihre Zuordnung zu einer konkreten Vernetzungsposition im Ausgangsartikel. Solange man keine Redundanzen in Kauf nehmen möchte, bedeutet aber auch hier die Auslagerung, dass aus den einzelnen Dokumenten die zugehörige HTML-Präsentation nicht vollständig ermittelt werden kann.

2.5 Zusammenfassung: Typen von Separabilität der Mediostruktur

Es hat sich gezeigt, dass die Möglichkeit, auf sinnvolle Weise mediostrukturelle Informationen aus den XML-Dokumenten der Wortartikel auszulagern, nur unter recht speziellen Bedingungen gegeben ist. Wenn man aus der Kenntnis von Bezugs- und Zieladresse einer Vernetzung mechanisch die konkrete Position der Vernetzungsangabe im XML-Dokument bestimmen kann, kann man dieses Adressenpaar, oben Vernetzungsspezifikation genannt, in einer eigenen Datenstruktur verwalten (Separabilität der Mediostruktur); ist die konkrete Position hingegen selber Gegenstand genuin lexikografischer Entscheidungen, kann allenfalls eine Kombination aus Vernetzungsspezifikation und positionellen Indizes ausgelagert werden (schwache Separabilität), was bereits zu einer wenig wünschenswerten doppelten Repräsentation von hierarchisch-positionalen Informationen führt. – Gibt es auf die Vernetzungen bezogene Angabezusätze, sind beide genannten Formen von Separabilität wenig relevant, da man sinnvollerweise die (allerdings möglicherweise selber komplexe XML-Teilbäume bildenden!) Angabezusätze mit auslagern können sollte; ist dies der Fall, sprechen wir hier von erweiterter bzw. erweiterter schwacher Separabilität – aber nur dann, wenn es keine auf mehrere Vernetzungen gleichzeitig bezogenen Zusätze gibt. Anderenfalls kommt es zwangsläufig zu einer doppelten Repräsentation der gesamten wortartikelinternen, durch die XML-Struktur vorgegebenen hierarchischen Gruppierung der Vernetzungsangaben. – In allen Fällen steht jedoch die Option zur Verfügung, nur die strukturellen Vernetzungsspezifikationen eines Artikels auszugliedern, indem man eineindeutige Adressen an die Vernetzungsangaben selbst vergibt.

3. Anwendungen und Schlussfolgerungen

3.1 Computerlexikografische Behandlung von Vernetzungen in *elexiko*

Die Vernetzungen in *elexiko*-Wortartikeln sind, wie gesehen, ein Beispiel für den Fall, dass die Mediostruktur eines Wörterbuchs aufgrund ihrer Komplexität allenfalls schwach separabel ist und die computerlexikografischen Kosten einer Ausgliederung ersichtlich höher als ihr Nutzen wären. Da die lexikografische Bearbeitung der XML-Instanzen von *elexiko* ursprünglich ausschließlich mit einem handelsüblichen XML-Editor durchgeführt wurde, war es sinnvoll zu fordern, dass die Vernetzungsadressenangaben lokal in den Instanzen erscheinen, so dass auch die Speicherung struktureller Vernetzungsspezifikationen außerhalb der XML-Instanzen keine Option war. Mittlerweile arbeitet das *elexiko*-Projekt mit einem im Haus entwickelten Editor-Plugin für das Vernetzungsmanagement (Meyer 2011).[11] Im Rahmen eines konservativen Herangehens wurde entschieden, die grundsätzlichen Datenstrukturen und die mit ihnen verbundenen Arbeitsabläufe nicht umzugestalten. Um die ein- und ausgehenden Vernetzungen eines im XML-Editor bearbeiteten Artikels aufzufinden sowie Integrität und Konsistenz seiner Vernetzungen zu prüfen, führt das Plugin folgende Prozesse durch:

- Das Dokument im Editor wird geparst und alle Vernetzungsangaben sowie die allgemeine Artikelstruktur (Lesarten, Adressen) werden ermittelt;

[11] Auch wenn grundsätzlich schon aus Wirtschaftlichkeitsgründen die Verwendung eines handelsüblichen Wörterbuch-Redaktionssystems wünschenswert ist, kann es doch – *pace* de Schryver (2011) – gute Gründe für eine gegenteilige Entscheidung geben. Im Fall von *elexiko* sind bereits die komplexe, schon vorhandene Hardware- und Software-Infrastruktur und die in lexikografischen Forschungsprojekten zu erwartenden häufigen Umstrukturierungen des XML-Schemas sowie die Existenz von Vernetzungen zwischen verschiedenen Wörterbüchern mit unterschiedlichen Schemata solche Gründe.

- alle Artikel, auf die sich die gefundenen Vernetzungsadressenangaben beziehen, werden, sofern wirklich vorhanden, aus der Datenbank geholt und geparst, um die Vernetzungsadresse zu prüfen;
- mit einer XPath-basierten Suche auf einem für XML optimierten Volltextindex werden alle Artikel in der Datenbank gesucht und ausgelesen, die eine Vernetzung auf das Dokument im Editor haben;
- diese Artikel werden ebenfalls geparst, um die zugehörigen Vernetzungsspezifikationen zu ermitteln;
- abschließend werden alle gefundenen Informationen miteinander abgeglichen und die Resultate in zwei Tabellen angezeigt, die für ein- bzw. ausgehende Vernetzungen die Vernetzungsspezifikationen und Konsistenzstatusinformationen liefern.

Aufgrund der relativ kleinen Menge an bearbeiteten Wortartikeln (< 2000) in *elexiko* ist dieses Vorgehen trotz der bemerkenswert komplexen Vernetzungsstruktur der Artikel und der hohen Kosten einer Volltextsuche für den lexikografischen Alltag hinreichend performant; eine Analyse der beschriebenen Art nimmt wenige Sekunden in Anspruch. Würde man ein skalierbareres Verfahren benötigen, wäre dennoch kein Auslagern von mediostruktureller Information erforderlich, da grundsätzlich noch ein anderes Verfahren zur Verfügung steht: Man kann die für Integritäts- und Konsistenzprüfungen benötigten Vernetzungsspezifikationen einfach in eine separate Tabelle *kopieren*, die über einen Datenbanktrigger bei jeder Änderung eines Artikels geprüft und ggf. aktualisiert wird. Die separate Tabelle fungiert dann als schneller relationaler Cache der benötigten Informationen (vgl. Joffe/Schryver/Prinsloo 2003; Meyer / Müller-Spitzer 2010).

3.2 Zusammenfassung

Vor dem Hintergrund einer relationalen lexikografischen Datenmodellierung schrieben Blumenthal et al. im Jahre 1988:

> Auf der Ebene der konzeptionellen Datenmodellierung gibt es Beziehungen. Beziehungen sind ungerichtet, d.h. in beiden Richtungen in gleicher Weise zugreifbar. Dies liegt daran, daß auf der konzeptionellen Ebene die Asymmetrie von Verweisursprung und Verweisziel im Hinblick auf ihre ‚Repräsentationsbedürftigkeit' verschwindet, da man dort direkt die Beziehung zwischen Verweisursprung und Verweisziel modelliert. (Blumenthal/Lemnitzer/Storrer 1988: 356)

Der vorliegende Beitrag ist ein Beleg dafür, dass durch die Verwendung von XML-basierten Repräsentationen die genannte Asymmetrie wieder in die Computerlexikografie Einzug gehalten hat; der „Verweisursprung" ist in den XML-Instanzen, wie auch auf der Präsentationsebene, „nicht symbolisch repräsentiert, sondern qua Lokalität … faktisch gegeben" (ebd.: 145). Der Versuch, Vernetzungen doch relational zu modellieren, führt, wie gezeigt, in den meisten Fällen zu unbrauchbaren Ergebnissen. Dies ist zunächst eine sehr unbefriedigende Diagnose, denn durch den Verzicht auf eine solche Modellierung kehren viele Probleme zurück, die man mit der Verwendung eines relationalen Datenbanksystems lösen konnte. Der einzige sinnvolle Ausweg ist eine Doppelstrategie, die es bei einer ausschließlich XML-basierten Datenhaltung und lokal in den Instanzen repräsentierten Vernetzungsadressenangaben belässt und mediostrukturelle Informationen in einem stets aktuell gehaltenen relationalen Repräsentationsformat vorhält, das performante Prüfungen und Suchvorgänge ermöglicht.

4. Literatur

Blumenthal, Andreas / Lemnitzer, Lothar / Storrer, Angelika (1988): Was ist eigentlich ein Verweis? Konzeptionelle Datenmodellierung als Voraussetzung computergestützter Verweisbehandlung. In: Harras, Gisela (Hg.): Das Wörterbuch. Artikel und Verweisstrukturen. (= Jahrbuch 1987 des Instituts für deutsche Sprache). Düsseldorf/Bielefeld, S. 351-373.

Haß, Ulrike (Hg.) (2005): Grundfragen der Elektronischen Lexikographie: *elexiko* – Das Online-Informationssystem zum deutschen Wortschatz. Berlin / New York.

Joffe, David / de Schryver, Gilles-Maurice / Prinsloo, Daniel Jacobus (2003): Computational features of the dictionary application "TshwaneLex". In: Southern African Linguistics and Applied Language Studies 21(4), S. 239-250.

Kammerer, Matthias (1998): Hypertextualisierung gedruckter Wörterbuchtexte: Verweisstrukturen und Hyperlinks. Eine Analyse anhand des FRÜHNEUHOCHDEUTSCHEN WÖRTERBUCHES. In: Storrer, Angelika / Harriehausen, Bettina (Hg.): Hypermedia für Lexikon und Grammatik. Tübingen, S. 145-171.

Klosa, Annette (Hg.) (2011): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur Deutschen Sprache 55). Tübingen.

Meyer, Peter (2011): vernetziko: a cross-reference management tool for the lexicographer's workbench. In: Kosem, Iztok / Kosem, Karmen (Hg.): Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex2011, Bled, Slowenien, 10-12 November 2011. Ljubljana, S. 191-198. Internet: http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-25.pdf. (Stand: Oktober 2012).

Meyer, Peter / Müller-Spitzer, Carolin (2010): Consistency of sense relations in a lexicographic context. In: Mititelu, Verginica Barbu / Pekar, Viktor / Barbu, Eduard (Hg.): Proceedings of the workshop "Semantic Relations. Theory and Applications", 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta. Internet: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf. (Stand: Oktober 2012).

Müller-Spitzer, Carolin (2007a): Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. In: Hermes 38, S. 137-171.

Müller-Spitzer, Carolin (2007b): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. (= Studien zur Deutschen Sprache 42). Tübingen.

Schryver, Gilles-Maurice de (2011): Why opting for a dedicated, professional, off-the-shelf dictionary writing system matters. In: Akasu, Kaoru / Uchida, Satoru (Hg.): ASIALEX 2011 Proceedings. LEXICOGRAPHY: Theoretical and practical perspectives. Papers submitted to the Seventh ASIALEX Biennial International Conference. Kyoto Terrsa, Kyoto, Japan, August 22-24, 2011. Kyoto, S. 647-656.

Wiegand, Herbert Ernst (1996): Über die Mediostrukturen bei gedruckten Wörterbüchern. In: Zettersten, Arne / Pedersen, Viggo Hjørnager (Hg.): Symposium on Lexicography VII. Proceedings of the Seventh International Symposium on Lexicography, May 5-6, 1994, at the University of Copenhagen. (= Lexicographica, Series Maior 76). Tübingen, S. 11-43.

Wiegand, Herbert Ernst (2002): Altes und Neues zur Mediostruktur in Printwörterbüchern. In: Lexicographica 18, S. 168-252.

Wilson, Mark (2006): Wandering significance. An essay on conceptual behaviour. Oxford / New York.